Probability of Expert Selection in Mixture of Experts (MoE) LLM Architecture

Problem Statement

Consider a Mixture of Experts (MoE) large language model (LLM) architecture with M total experts. For each token prediction, exactly k experts are randomly chosen from the M available experts. This selection is performed independently for each of the n tokens in a sequence.

We aim to find the probability that a particular expert is selected *at least once* during the sequence of n token predictions.

Solution

To solve this, we can break down the problem into the following steps:

1. **Calculate Probability of Non-selection per Token (ref to next page for derivation):** For each token, the probability that a specific expert is *not* selected among the k chosen experts is:

$$\frac{M-k}{M}$$

2. **Probability of Non-selection across n Tokens:** Since each token's selection is independent, the probability that the expert is *never* selected across n tokens is:

$$\left(\frac{M-k}{M}\right)^n$$

3. **Probability of Selection at Least Once:** The probability that the expert is selected *at least once* during the sequence of n tokens is the complement of the non-selection probability:

$$P(\text{selected at least once}) = 1 - \left(\frac{M-k}{M}\right)^n$$

Final Answer

Thus, the probability that a particular expert is selected at least once during n token predictions is:

$$1 - \left(\frac{M-k}{M}\right)^n$$

Derivation: Calculate Probability of Non-selection per Token

Consider a Mixture of Experts (MoE) architecture comprising M experts. For each token prediction, exactly k experts are randomly chosen from the M available experts. We aim to find the probability that a specific expert (say, Expert A) is **not** selected among the k experts chosen.

1 Total Possible Selections

Each time we choose k experts out of M, the total number of possible ways to make this selection is given by:

$$\binom{M}{k} = \frac{M!}{k!(M-k)!}$$

2 Number of Favorable Outcomes

To calculate the probability of "Expert A" **not** being selected, we count the ways to choose k experts from the remaining $M-1$ experts (excluding Expert A):

$$\binom{M-1}{k} = \frac{(M-1)!}{k!(M-1-k)!}$$

3 Calculating the Probability

The probability that "Expert A" is not selected among the k chosen experts is given by the ratio of favorable outcomes to total possible selections:

$$P(\text{Expert A is not selected}) = \frac{\binom{M-1}{k}}{\binom{M}{k}}$$

4 Simplifying the Expression

Substituting the values of the combinations:

$$P(\text{Expert A is not selected}) = \frac{\frac{(M-1)!}{k!(M-1-k)!}}{\frac{M!}{k!(M-k)!}}$$

Canceling out $k!$ in both the numerator and the denominator:

$$= \frac{(M-1)!(M-k)!}{M!(M-1-k)!}$$

Since $M! = M \cdot (M-1)!$, we can rewrite this as:

$$= \frac{M-k}{M}$$